A Vessel Cost Sampling Model

John M. Gates[a] and Andrew Kitts[b]

ABSTRACT

The Social Sciences Branch (SSB) of NOAA's Northeast Fisheries Science Center (NEFSC) uses economic data to estimate the economic impacts of fisheries regulatory actions. Since 1995, the Northeast Fisheries Observer Program has collected trip cost information from commercial fishing vessels on which observers have been deployed. However, the allocation of observer coverage to fisheries is largely determined by non-economic needs (e.g., regulatory compliance and monitoring, bycatch estimation, and assessment of protected species interactions). To obtain cost data from a much wide variety of fisheries, the SSB funds additional observer coverage to acquire economic data from under-represented fisheries.

In this paper, we describe a mathematical programming model to allocate marginal observer coverage, in terms of observer sea days. We apply a simple logarithmic utility function to the fisheries of New England. The potential universe of observations is divided into strata based on vessel size, gear type, and state of landing. Certain *a priori* assumptions underlie the approach:

- Costs are assumed heterogeneous across strata until such time as statistical analysis supports the null hypothesis
- The marginal utility of observations in a given cell is subject to diminishing returns
- The utility of observations decays with age
- The marginal utility of zero observations is positive and finite
- The cost-effectiveness of data collection can be reduced by software that automates the process

In addition:

- Data collection each year is limited by a budget
- Current data collection takes into account the amount of past data for each cell and most data will be collected via other programs

INTRODUCTION

We describe an approach to designing a sampling program for the collection of trip cost information from commercial fishing vessels operating off the northeastern coast of the United States. The Northeast Fisheries Observer Program uses at-sea observers to collect both biological and economic information from federally permitted fishing vessels. Deployment of observers to specific fleet components is driven by a wide variety of management needs. Most observer coverage is allocated for estimating bycatch, evaluating protected species interactions, and monitoring regulatory compliance. While economic information is collected on these trips, the distribution of observer days by vessel type and geographic region is not determined by analysts using the economic data. However, as funds are available, the Social

Sciences Branch (SSB) of NOAA's Northeast Fisheries Science Center (NEFSC) allocates observer days to fisheries not normally sampled. The sampling model described here was developed as a tool for optimally allocating SSB funded observer coverage to fleet components under-represented in the Observer Program database.

In principle, a stratified random sampling design would seem obvious. However, a truly random sample is an unattainable ideal as various data users sample fleet components using different objectives. Instead, we focus on defining the critical components of a supplemental sampling program, and then build these components into a utility function for any incremental observer coverage acquired using SSB funding. An idealized utility function would have some of the following attributes defined over the potential universe of observations:

- Costs are assumed heterogeneous across strata until such time as statistical analysis supports the null hypothesis

- The marginal utility of observations in a given cell is subject to diminishing returns. This assumption seems reasonable since, from a statistical perspective, confidence intervals decline with sample size.

- The utility of observations decays with age. The contrary assumption implies that once there are enough observations in a cell, no more data need be collected. However, a time series of vessel cost data is desired in order to keep the data current. This case can be represented by a decay rate of zero.

- The marginal utility of zero observations is positive and finite

- The cost-effectiveness of data collection can be reduced by software that automates the process

In addition:

- Data collection each year is limited by a budget

- Current data collection takes into account the amount of past data for each cell and that most data will be collected via other programs

DESCRIPTION OF COST DATA COLLECTION PROGRAM

Cost information is needed by fishery economists to further understand the factors that influence business decisions made by vessel owners. The data are also used to estimate the economic impact of proposed fishing regulations (trip limits, closed areas, days-at-sea, quotas, etc.) and how these regulations are likely to affect different industry components, different ports, and different groups of fishermen and fishing communities. With better data about costs, fishery managers should be able to better develop regulations which increase the net benefits derived from fishery resources.

To undertake such analyses, information is needed on both annual costs and trip costs. Annual costs include insurance, mooring and dockage fees, repairs and maintenance, and other expenses not directly determined by the number or duration of trips made by a fishing vessel. Trip costs include ice, fuel, bait, food for the crew, etc.

NOAA's fishery observers currently collect information on trip costs. Annual costs are collected by a separate mail survey of vessel owners. Fishing vessel cost data have been collected through the Observer Program since 1995. However, until 2003, the focus was only on a few gear types. Since 2003, largely due to increases in bycatch/discard monitoring, the number of observer days has markedly increased (from around 2,000 days in the late 1990s/early 2000s to over 4,000 days in 2003, and over 8,000 days in 2005). The 2008 schedule calls for 4,000 observer days.

The Observer Program is a convenient way to collect cost data on a continuous basis from a variety of fisheries. Little time is required to record the cost information and since observers are already onboard the vessel to primarily collect biological information, it is a practical way to obtain economic information with little additional expense. Collecting this information while the trip is underway is also less burdensome to fishing captains and crew as compared to involving them in a land-based survey.

MODEL DESCRIPTION

The Utility Function

A log-linear utility function was used in the sampling model. Abstracting from constraints and the fine details of the strata, a typical term in the utility function would have the form U:

$U = \text{Log}(1 + \text{newn} + \text{exogn} + \Sigma_a \rho(a) * \text{oldn}(a))$
where:

newn = number of supplemental observer days to sample within a stratum
exogn = number of observer days that will be sampled by other programs within a stratum
oldn(a) = number of existing observer days within a stratum of age a
$\rho(a)$ = vintage discount factor, e^{-ra} where r = discount factor applied to age of data; $a \geq 0$

Such a utility function tends towards a uniform distribution of observations across strata, but the model adjusts (a) for the age of existing data and (b) for exogenous data to be collected under other programs. Accounting rows were also defined to generate information of interest. For example, a pro-forma charge for the opportunity cost of the captain's time spent answering questions. These are aggregated and reported in the output.

To implement the model, GAMS software for mathematical programming[c] was used. The model presented below is non-linear, but it is easily approximated to any desired degree of precision using grid-linearization. Such linearization techniques allow use of linear programming (LP), rather than non-linear programming (NLP). One caveat of the model is that it is a normative guide to be used adaptively. That is, it is intended to help staff automate the process of selecting supplemental observer days subject to a budget constraint.

Strata

Exploration of existing trip cost data suggests that costs are determined primarily by the type of gear used, vessel size, and geographic region. Therefore, the data were stratified around gear type, three vessel sizes (small, medium, and large), and state of landing. Since costs change over time as both technology and the price of inputs change (for example, fuel prices increase), data vintage is also an important component. In GAMS terminology, the strata and the time component were defined as sets:

G = {g} = gear indices
S = {s} = vessel size indices
ST ={st} = state indices
Y = {y} = year index

Parameters

The GAMS model uses quite simple parameters that are either logical or based on historical patterns of fishing effort and data collected. These include the size of populations to be surveyed.

budget = The budget limit.

chrtcost(G,S,ST) = Unit cost of chartering an observer day

exogn(G,S,ST) = For the year of analysis, a table of the number of observer days to be collected by non-SSB programs

maxfract is a subset of [0,1] = Maximum expected participation rate

ncost(G,S,ST) = Pro-forma unit charge to reflect the opportunity cost per question of the respondent[d]

oldn(G,S,ST,Y) = The number of observer days, previously collected from each stratum in all years prior to the year of analysis

tpop(G,S,ST) = For the year of analysis, a table of the number of observer days, by gear type, vessel size, and state of landing, of the total population of fishing vessel activity in the Northeast. This provides an upper bound on the number of observer days that can be allocated to a stratum.

Primal Variables

Math programming (MP) models involve two types of decision variables. The first type is called primal (or column) variables. The second type is called dual (or row) variables. Every MP problem has both a primal and associated dual problem. The choice of labels “primal problem” and “dual problem” is a matter of convenience for the user, but the labeling of primal and dual variables switches when the problems are switched. In the present case, utility maximization was selected as the primal problem. Consequently, the primal variables of principal interest are the number of trips sampled by stratum. The dual variables indicate how much the solution would change if a variable were increased or a constraint equation relaxed. This point will be returned to later.

Non-Negative Variables

In statistical estimation, it is not (usually) necessary to specify the sign of variables. Negative variables are not usually observed (it depends on their definition). However, in MP the distinction must always be explicit. Some variables can be negative or positive depending on how the equations are defined.

IMPTCOST = Imputed cost of captains' time answering questions
NCHARTERS(G,S,ST) = Planned observer days chartered (days per year^-1)
TCHARTERS = Aggregate of days to be surveyed
NEWN(G,S,ST) = This is the output from the model. It is a table of the number of supplemental observer days that should be allocated to each stratum.

NSAMPLE = Total observations in the new sample
SURCOST = Survey cost

Free Variables

MP problems typically have at least one free variable that measures the value of the objective function. Solution software typically requires that this variable be unrestricted. However, sometimes an analyst wishes to know certain items for reporting purposes. These items involve free variables and they do not affect the solution.

UTILITY = Utility of additional observer days; this is what the model maximizes.

Bounded Variables
The default bounds on primals are (0, ∞). Frequently, *a priori* knowledge exists that allows tighter bounds. This is achieved by explicit bounds. Any variable can be bounded by adding an explicit constraint equation for each bound. However, this is an inefficient way to explicitly bound variables, because it requires much more coding. GAMS allows one to define bounds on variables with one-line entries in a BOUNDS section.

NCHARTERS(G,S,ST). maxfract*tpop(G,S,ST)
This provides an upper bound on NCHARTERS(.). In conjunction with RNEWN(.) below; it is also an upper bound for NEWN(.).

SURCOST. budget
This is the budget limit.

Equations

Historically, the earliest optimization problems were unconstrained. Later, optimization problems with equality constraints were solved. Optimization with inequality constraints remained an unsolved problem until LP was invented.[e] Shortly thereafter, Kuhn and Tucker offered a theorem for linear and non-linear programming, and thus MP was created. As a practical (numerical solutions) matter, many continuous time control theory problems were first solved using discrete-time mathematical programming and many problems are still solved in this manner. So, it is a fair generalization to assert that most MP problems involve constraints. At some level of abstraction and policy analysis, all relevant constraints can be considered to be imbedded implicitly in data and that therefore deemed as superfluous. This is not the operational context addressed in this paper.

$$\text{RUtility.. UTILITY} \leq \Sigma_{\text{G,S,ST}}\ \text{LOG}(1+\Sigma_{\text{Y}}\ \rho(\text{Y,G})*\text{oldn(G,S,ST,Y)}$$

$$+ \text{exogn(G,S,ST)} + \text{NEWN(G,S,ST))\$(tpop(G,S,ST)} > 0))$$

This equation defines the objective function to be optimized. The expression $(.) is the GAMS analogue of a logical IF statement which causes the algebra to which it applies to be switched on if the bracketed statement is true; zero otherwise. Thus, the utility index applies only to non-empty strata (many strata are empty).

RNEWN(G,S,ST).. NCHARTERS(G,S,ST). NEWN(G,S,ST)
A trip charter for a stratum is a necessary condition for new data to be collected from that stratum.

RTCHARTERS.. TCHARTERS $\geq \Sigma_{G,S,ST}$ NCHARTERS(G, S, ST))
This is an accounting equation that aggregates charters across strata.

RNSAMPLE.. - $\Sigma_{G,S,ST}$ NEWN(G,S,ST)) + NSAMPLE = 0
This is an accounting equation that aggregates observations across strata

RNPTRIP.. NPERTRIP $\geq$ (NSAMPLE/TCHARTERS); TCHARTERS > 0; 0 otherwise
This is an accounting equation that calculates the mean questions asked per trip.

RTCOST.. IMPTCOST $\geq \Sigma_{G,S,ST}$ ncost(G,S,ST)*NEWN(G,S,ST))
This equation calculates a pro-forma estimate of the time costs imposed on the captain in answering the questions; ncost(.) $\geq$ 0.

RSURVCOST.. SURCOST. chrtcost(G,S,ST)*TCHARTERS(G,S,ST) + IMPTCOST
Survey costs include trip chartering costs plus the imputed costs of the captain's time.

RESULTS

Tables 1 through 3 show the data inputs used in the GAMS model to allocate observer days in 2006. Table I is the number of observer days deployed in 2001 through 2005, by state and gear (oldn – vessel size stratum not shown). Table II is the number of observer days scheduled to be deployed in 2006 by non-SSB NEFSC departments (exogn – vessel size stratum not shown). Table III is the total number of actual fishing days taken by federally permitted vessels in 2005 (tpop – vessel size stratum not shown). The parameter tpop provides a cap so that allocated observer days do not exceed actual effort within a cell. In 2006, the budget limit was $50,000 and the cost per observer day way $800. Therefore, the model allocates 62.5 days across strata.

Table I. Observer Days Deployed in 2001 through 2005 by Gear and State (oldn – vessel size stratum not shown)

	Year	CT	DE	MA	MD	ME-NH	NC	NJ	NY	RI	VA	Total
Scallop Dredge	2001	14		108				41		9	50	222
	2002	17		122				123			257	519
	2003	3		340				257			436	1,036
	2004			470				195			156	821
	2005			259								259
Sink Gillnet	2001	2	2	369	9	46	90	4	28	56	322	928
	2002	5		411	8	32	16	15	6	68	181	742
	2003			566		33	37	13	13	71	86	819
	2004			988		70	111	97	29	74	99	1,468
	2005			741		142	92	94	6	66	117	1,258

Table I (continued). Observer Days Deployed in 2001 through 2005 by Gear and State (oldn – vessel size stratum not shown)

Bottom Longline	2001			37		6			11			54
	2002			16		5			21			42
	2003			127								127
	2004			363								363
	2005											
Otter Trawl, Scallop	2001							23	7		54	84
	2002							72	9		181	262
	2003										6	6
	2004											
	2005											
Otter Trawl, Fish	2001	10		417	48	81	29	118	47	76	32	858
	2002	42		868	41	142	43	95	48	134	50	1,463
	2003	17		1,286	8	395	103	155	6	239	38	2,247
	2004	8		1,708	10	399	107	304	69	466	46	3,117
	2005	15		5,569		600		137	101	530		6,952
Midwater Trawl	2001			9		4		31		11		55
	2002			14		6		8				28
	2003			54		35						89
	2004					54						54
	2005											
Otter Trawl, Shrimp	2001	6		8		8		3				25
	2002			2		3						5
	2003			2		11						13
	2004					12						12
	2005											
Fish Pot	2001			5	5		1			6		17
	2002									1		1
	2003											
	2004											
	2005											
Lobster Pot	2001			20						4		24
	2002											
	2003											
	2004											
	2005											
Pair Trawl	2001			46		13		12		6		77
	2002			104		74		11		11		200
	2003			176		127		6				309
	2004											
	2005											

Table I (continued). Observer Days Deployed in 2001 through 2005 by Gear and State (oldn – vessel size stratum not shown)

Purse Seine	2001			6		45						51
	2002					75						75
	2003											
	2004											
	2005											
Scottish Seine	2001			5								5
	2002			5								5
	2003			8								8
	2004			8								8
	2005											

Table II. Scheduled Observer Days by Non-SSB Departments in 2006 (exogn – vessel size stratum not shown)

	CT	DE	MA	MD	ME-NH	NC	NJ	NY	RI	VA	Total
Scallop Dredge	3		266				351			451	1,071
Sink Gillnet	2	0	845	42	426	549	641	127	144	285	3,061
Bottom Longline			79				0	38			117
Otter Trawl, Scallop							129			192	321
Otter Trawl, Fish	100		893	17	256	154	502	349	1283	92	3,646
Midwater Trawl			32		57		50		4		143
Otter Trawl, Shrimp			2		10						12
Pair Trawl			180		105		16		8		309
Purse Seine			0		100						100
Total	105	0	2,297	59	954	703	1,689	514	1,439	1,020	8,780

Table III. Total Number of Actual Days Fished in 2005 (tpop – vessel size stratum not shown)

	CT	DE	MA	MD	ME-NH	NC	NJ	NY	RI	VA	Total
Clam Dredge			315	399			1,246	70	5		2,035
Scallop Dredge	1,142	78	18,822	1,214	266	75	12,779	764	1,667	7,069	43,876
Drift Gillnet		25	2	17		98	151	105		117	515
Sink Gillnet	168	2	6,620	145	3,513	646	2,242	2,230	1,822	849	18,237
Bottom Longline			1,604		188	11	155	397	3		2,358
Pelagic Longline			3	4		26	63				96
Otter Trawl, Scallop			57	266		47	1,580	68	8	2,350	4,376
Otter Trawl, Fish	1,995		21,995	664	6,424	2,920	5,815	8,449	9,601	3,904	61,767
Midwater Trawl			247		272	14	100		140	10	783
Otter Trawl, Fish			73		2,267	149	14		1	159	2,663
Crab Pot			528		20	24	54	556	68		1,250
Fish Pot	6	63	636	250	9	10	469	742	661	163	3,009
Lobster Pot	830	11	15,515	69	17,727		1,064	2,610	4,960	72	42,858
Pair Trawl			1,107		467		225		63		1,862
Scottish Seine			20					5			25
Total	4,141	179	67,544	3,028	31,153	4,020	25,957	15,996	18,999	14,693	185,710

Table IV summarizes the results from applying the data inputs to the GAMS (the total in Table IV is 63.4 days due to rounding). The output is the number of observer days to be allocated by stratum – state of landing (set ST), gear type (set G), and vessel size (set S). However, for ease of presentation and discussion, the vessel size stratum is not shown in Table IV and the results are aggregated over state and gear. The model allocates days in partial days so the end user must round to whole-day units. As indicated in Table IV, with a budget of $50,000 for additional observer coverage, the model spreads the observer days over a wide variety of under-observed cells. At higher budget levels, these cells would reach levels of maximum utility at different points and the results would show a range of values rather than identical values in each cell.

Table IV. GAMS Model Results – Observer Day Allocation by Gear and State (vessel size stratum not shown)

	CT	DE	MA	MD	ME-NH	NC	NJ	NY	RI	VA	Total
Clam Dredge			1.2	1.2			1.2	1.2	1.2		6
Scallop Dredge		1.2		1.2	1.2	1.2		1.2			6
Drift Gillnet		1.2	1.2	1.2		1.2	1.2	1.2		1.2	8.4
Bottom Longline						1.2	1.2		1.2		3.6
Pelagic Longline			1.2	1.2		1.2	1.2				4.8
Otter Trawl, Scallop			1.2	1.2		1.2			1.2		4.8
Midwater Trawl						1.2				1.2	2.4
Otter Trawl, Shrimp						1.2			1	1.2	3.4
Crab Pot			1.2		1.2	1.2	1.2	1.2	1.2		7.2
Fish Pot	1.2	1.2			1.2		1.2	1.2		1.2	7.2
Lobster Pot			1.2	1.2	1.2		1.2	1.2	1.2	1.2	8.4
Scottish Seine								1.2			1.2
Total	1.2	3.6	7.2	7.2	4.8	9.6	8.4	8.4	7	6	63.4

The GAMS model is easy to update each year with new inputs. In so doing, considerations such as the distribution and vintages of existing observations are taken into account. Of course, for various reasons, it may not prove possible to actually obtain observations in some cells. In such a case, the model output provides a guide to those cells next favored as substitutes for the preferred cells. It is expected that the results are not unique and so making substitutions would entail little or no reduction in the objective function.

The Kuhn-Tucker conditions have been omitted but these can be provided upon request. One insight that emerged from considering these conditions is what to do if a planned trip is not possible for whatever reason. A sensible way forward would be to draw another trip from those trips for which the primal values are zero (non-basic) and for which the dual or marginal values are zero or only slightly negative (the marginals [or duals] for primal variables measure the marginal gain from bringing a variable into the "solution"). Generally, if a primal variable is positive, the associated marginal, or dual, is zero. For non-basic primals (those with zero value), all the dual or marginals are negative -- if the optimal solution is unique. The closer to a multiple near optima solution, the closer some of the marginals will be to zero. If there are multiple true optima, some zero primals can be brought into the solution with <u>no</u> change in the objective function. These constitute a good sample set to use for substitutions should a planned observer trip not be possible. Multiple near optima are most likely in early years since many strata have zero data and hence the gradients of the objective function are all equal for these empty cells. Which of these equivalent strata is then selected for observer coverage is arbitrary, unless differences exist in the costs of charters or the opportunity costs of the respondents.

DISCUSSION

The GAMS model is a helpful tool for automating the decision making process regarding the allocation of supplemental observer days funded by the SSB. As written, the model treats each cell independently of all other cells. This provides a basis for quickly determining which cells are under-represented and how additional days might best be allocated. However, in practice each cell is not completely independent. That is, the cost of operating a large scallop dredge vessel in Massachusetts may be similar to that of a large scallop dredge vessel in New Jersey. Similarly, a small drift gillnet vessel may have a similar cost structure to a small sink gillnet vessel. A useful extension to the model would be to allow the user to place different levels of priority, based on an examination of the cost data, on each stratum. For example, the gear type stratum could be given higher priority over vessel size and, lastly, state. This way, the model could first allocate days based on the relative presence/absence (and vintage) of observations in the different gear categories before allocating those days across vessel size and state.

Some of this flexibility can be achieved with the current model by aggregating some of the strata. For example, rather than using each individual state, states can be combined into New England and Mid-Atlantic regions. Gear types found to have similar cost can be similarly aggregated.

ENDNOTES

[a] Department of Environmental and Natural Resource Economics, University of Rhode Island

[b] NOAA Fisheries, Northeast Fisheries Science Center , Woods Hole, MA

[c] GAMS is an acronym for General Algebraic Modeling System. It is a high level programming language written and compiled in FORTRAN. It allows a user to express a model in algebra-like statements. From GAMS one can invoke any of the major commercial Solvers to optimize the problem.

[d] We developed the model to allow for different sampling environments. In one environment, the number of questions asked of a respondent is not fixed and we wanted to allow for the opportunity cost of the respondents time. In this case, an additional set, which represents all the survey questions, can be defined in the model. When a trip is surveyed by on-board observers, all questions (trip costs) are collected at the same time. In such a case, the set can be collapsed to one entry even though multiple questions will be asked. In such a case, the respondent's opportunity cost (ncost(.)) can be rescaled to a per questionnaire rather than a per question basis. In the interests of simplicity, we have omitted the set C = { a set of questions} to be asked. It is a one-line change to revert to the more general case if it is relevant.

[e] As a matter of history, LP can be said to have been invented in the late 1920s by the Russian mathematician Kantorovich. He offered it as a "..method to improve the allocation of resources in a Socialist economy".